การเปรียบเทียบการประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธีกำลังสองน้อยที่สุด วิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของริดจ์และสไตน์ ในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ

(A COMPARISON AMONG ORDINARY LEAST SQUARES, RIDGE REGRESSION, AND RIDGE AND STEIN METHODS IN ESTIMATING MULTIPLE REGRESSION WITH MULTICOLLINEARITY)

ธันยากร ต้นชลขันธ์[1]

1. ความเป็นมาและความสำคัญของปัญหา

ในการศึกษาการเปลี่ยนแปลงของตัวแปรตาม (dependent variable) ว่ามีผลมาจากตัวแปรอิสระ (independent variable) ชุดหนึ่งอย่างไรนั้น วิธีการหนึ่งที่ใช้ในการหารูปแบบความสัมพันธ์ของตัวแปรเหล่านั้นคือ การวิเคราะห์ความถดถอยพหุ (multiple regression analysis) ซึ่งเป็นกรณีหนึ่งของการวิเคราะห์ความถดถอยเชิงเส้น ในการวิเคราะห์ความถดถอยพหุมีหลักเกณฑ์ว่า การใช้ตัวแปรอิสระที่เหมาะสมมากกว่าหนึ่งตัวโดยทั่วไปย่อมทำให้ผลการประมาณค่าตัวแปรตามมีความถูกต้องมากกว่าการใช้ตัวแปรอิสระเพียงตัวเดียวถ้าตัวแปรอิสระนั้นไม่มีความสัมพันธ์กันและมีอิทธิพลต่อตัวแปรตามมากพอสมควร เราสามารถเขียนตัวแบบทั่วไป (general model) ของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเชิงเส้นได้ดังนี้

(1) $$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x1

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด n x (p+1)

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยพหุขนาด (p+1) x 1

$\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของค่าความคลาดเคลื่อนที่เกิดขึ้นขนาด n x 1

โดยที่ $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$, $\text{cov}(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

---

[1] นักศึกษาปริญญาโท สาขาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุจากตัวแบบดังกล่าวนี้ วิธีที่นิยมใช้กันมากที่สุด คือ วิธีกำลังสองน้อยที่สุด (Least squares method) โดยจะได้รูปแบบตัวประมาณเป็น $\hat{\beta} = (X'X)^{-1}X'y$ ซึ่งมีคุณสมบัติคือ เป็นตัวประมาณที่ไม่เอนเอียงและให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุดในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้น แต่ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดมีข้อสมมติที่จำเป็นข้อหนึ่ง คือ ตัวแปรอิสระจะต้องไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ซึ่งในทางปฏิบัติเป็นไปได้น้อยมากเพราะตัวแปรต่างๆ ที่นำมาศึกษาอาจมีความสัมพันธ์กัน กล่าวคือ ตัวแปรอิสระบางตัวมีพหุสัมพันธ์ (Multi-collinearity) กันทำให้การประมาณค่าตัวแปรตามที่ได้ไม่เหมาะสม และมีผลทำให้ค่าเฉลี่ยความคลาดเคลื่อกำลังสองของค่าสัมประสิทธิ์การถดถอยพหุมีค่ามากขึ้น นั่นคือ ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่ได้ขาดความเที่ยงตรง

Hoerl and Kennard (1970:55-67) ได้ศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด โดยให้ชื่อว่า วิธีริดจ์รีเกรสชัน (Ridge regression method) ซึ่งวิธีนี้ไม่ต้องตัดตัวแปรอิสระออกจากตัวแบบ หลักการของวิธีนี้คือ จะให้ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำลงโดยจะบวกค่าคงที่ค่าหนึ่งที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุมของเมตริกซ์ $X'X$ เนื่องจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธีกำลังสองน้อยที่สุดเป็นฟังก์ชันของ $(X'X)^{-1}$ ดังนั้นการที่จะพยายามลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองจึงต้องพยายามลดค่า $(X'X)^{-1}$ ให้ต่ำลงซึ่งจะทำได้โดยการบวกค่าคงที่ที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุม จะทำให้ได้ตัวประมาณสัมประสิทธิ์การถดถอยพหุด้วยวิธีริดจ์รีเกรสชัน คือ

$$\hat{\beta}_R = (X'X + kI)^{-1} X'y \qquad ; \quad k > 0$$

ตัวประมาณสัมประสิทธิ์การถดถอยพหุด้วยวิธีนี้มีคุณสมบัติคือ มีความเอนเอียง(bias) และจะต้องมีการประมาณค่าพารามิเตอร์ k ซึ่งได้มีผู้เสนอการประมาณค่าพารามิเตอร์ k หลายวิธี เช่น วิธีของ Hoerl Kennard and Baldwin (HKB) วิธีของ TZE-SAN-LEE (TZE) วิธีของ McDonald Galarneau(MC & D) เป็นต้น แต่ยังไม่มีการสรุปแน่นอนว่าวิธีการใดดีที่สุด จากปัญหานี้ได้มีผู้เสนอตัวประมาณใหม่ขึ้นมาโดยอาศัยหลักการของตัวประมาณริดจ์และอาศัยรูปแบบของตัวประมาณอีกตัวประมาณหนึ่ง คือ ตัวประมาณสไตน์ (Stein estimator) เนื่องจากรูปแบบของตัวประมาณดังกล่าวสามารถหาค่าพารามิเตอร์ได้ง่าย ซึ่งรูปแบบของตัวประมาณ คือ

$$\hat{\beta}_S = c\hat{\beta} \qquad ; \quad 0 < c < 1$$

Liu Keijian(1993:393-403) ได้เสนอตัวประมาณใหม่ขึ้นมาโดยนำเอาคุณสมบัติข้อดีของทั้งตัวประมาณริดจ์และตัวประมาณสไตน์มาใช้และมีรูปแบบของตัวประมาณ คือ

$$\hat{\beta}_c = (X'X+I)^{-1}(X'y+c\hat{\beta}) \quad ; \quad 0 < c < 1$$

ตัวประมาณใหม่นี้ได้ใช้หลักการของตัวประมาณริดจ์คือ การบวกค่าคงที่ค่าหนึ่งเข้ากับสมาชิกในแนวทแยงมุมของเมตริกซ์ $X'X$ ซึ่งค่าคงที่นั้นมีค่าเท่ากับหนึ่ง และใช้รูปแบบของตัวประมาณสไตน์บวกเข้าไปในเทอมหลังเพื่อจะแก้ปัญหาการเกิดพหุสัมพันธ์ให้กับเทอมหลัง เราสามารถพิจารณาได้ชัดเจนเมื่อจัดรูปแบบของตัวประมาณใหม่ โดยใช้สมการปกติของการวิ-เคราะห์วิธีกำลังสองน้อยที่สุดซึ่งมีรูปแบบคือ $X'X\hat{\beta} = X'y$ จะทำให้ได้รูปแบบของตัวประมาณดังสมการข้างล่างนี้

$$\hat{\beta}_c = (X'X+I)^{-1}(X'X\hat{\beta}+c\hat{\beta})$$

$$= (X'X+I)^{-1}(X'X+cI)\hat{\beta} \quad ; \quad 0 < c < 1$$

เมื่อพิจารณาโดยใช้ค่าเจาะจงของเมตริกซ์ $X'X$ ซึ่งมีคุณสมบัติข้อหนึ่งของค่าเจาะจงของเมตริกซ์ $X'X$ กล่าวคือถ้า λ_i เป็นค่าเจาะจงของเมตริกซ์ $X'X$ เมื่อ i = 1, 2, 3,..., p แล้ว $\sum_{i=1}^{p}\lambda_i = \text{trace}(X'X)$ ดังนั้นเราจะสมมติให้ค่าเจาะจงของเมตริกซ์ $X'X$ มีค่าเท่ากับ λ_i ซึ่ง $\lambda_{max} = \lambda_1$ เป็นค่าเจาะจงที่มีค่ามากที่สุด และ $\lambda_2 \geq \lambda_3 \geq ... \geq (\lambda_p = \lambda_{min})$ โดยที่ค่า λ_{min} เป็นค่าเจาะจงที่มีค่าน้อยที่สุดซึ่งมีค่ามากกว่าศูนย์

ถ้าเกิดปัญหา Multicollinearity จะทำให้เมตริกซ์ $X'X$ มีความเป็นเมตริกซ์เอกฐาน (singularity matrix) มากขึ้น กล่าวคือ ค่าเจาะจงมีค่าน้อย ผลที่ติดตามมาคือ λ_{min} มีค่าเข้าใกล้ศูนย์ จะได้ λ_{min}^{-1} มีค่าเข้าใกล้อนันต์ ตัวประมาณริดจ์ได้แก้ปัญหาในการเป็นเอกฐานโดยการบวกค่าคงที่ให้ λ_i แต่ละตัวเท่าๆ กัน ซึ่งคือค่าในเทอมแรก แต่เมื่อพิจารณาในเทอมหลังพบว่าอาจยังมีปัญหาอยู่เนื่องจากมีเทอมที่เกี่ยวข้องกับ $X'X$ ดังนั้นตัวประมาณที่ใช้หลักการของริดจ์และสไตน์ได้แก้ปัญหาโดยการบวกค่าคงที่ให้กับสมาชิกในแนวทแยงมุมทุกตัว ซึ่งมีค่าเท่ากับ c จึงทำให้ได้รูปแบบของตัวประมาณดังกล่าว และจากหลักการที่นำมาสร้างตัวประมาณใหม่ดังกล่าว เพื่อความสะดวกในการวิจัยผู้วิจัยจึงเรียกตัวประมาณใหม่นี้ว่า **ตัวประมาณที่ใช้หลักการของริดจ์และสไตน์ (Ridge and Stein estimator)** ซึ่งคุณสมบัติของตัวประมาณนี้คือ เป็นตัวประมาณที่เอนเอียงและให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุต่ำกว่าวิธีกำลังสองน้อยที่สุดเช่นเดียวกัน

จากคุณสมบัติในการประมาณค่าสัมประสิทธิ์การถดถอยพหุทั้ง 3 วิธีข้างต้น จึงเป็นสิ่งที่น่าสนใจว่า เมื่อตัวแปรอิสระมีความสัมพันธ์กัน การประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด วิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของริดจ์และสไตน์ วิธีใดจะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำที่สุด

2. วัตถุประสงค์ของการวิจัย

การวิจัยนี้มีวัตถุประสงค์ที่จะเปรียบเทียบการประมาณค่าสัมประสิทธิ์การถดถอยพหุ เมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยวิธีกำลังสองน้อยที่สุด วิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของริดจ์และสไตน์ เมื่อการแจกแจงของความคลาดเคลื่อนเป็นปกติ ปกติปลอมปน และแบบเบ้

3. สมมติฐานของการวิจัย

เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ ปกติปลอมปน และแบบเบ้ และตัวแปรอิสระมีพหุสัมพันธ์กันมากๆ วิธีริดจ์รีเกรสชัน วิธีที่ใช้หลักการของริดจ์และสไตน์จะให้ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่มีความถูกต้องมากกว่าวิธีกำลังสองน้อยที่สุดภายใต้ขนาดตัวอย่าง จำนวนตัวแปรอิสระ ระดับของสัมประสิทธิ์การแปรผันและชุดของสัมประสิทธิ์การถดถอยโดยใช้ข้อกำหนดเดียวกัน

4. ขอบเขตของการวิจัย

4.1 ความคลาดเคลื่อนมีการแจกแจงปกติ (Normal Distribution)

การแจกแจงความผิดพลาดเป็นไปตามข้อตกลงของสมการถดถอย คือ $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{1}, \sigma^2 I_n)$ ซึ่งฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \qquad \sigma > 0$$

ก) จำนวนตัวแปรอิสระที่ใช้ในการศึกษามี 2 ระดับ คือ 3 และ 5

ข) ขนาดตัวอย่างที่ศึกษามี 3 ขนาด คือ 30 , 50 และ 100

ค) ระดับของสัมประสิทธิ์การแปรผัน(C.V.) ที่ศึกษามี 3 ระดับ คือ 5%, 10% และ 15%

ง) ระดับพหุสัมพันธ์ของตัวแปรอิสระ

1) กรณีตัวแปรอิสระเท่ากับ 3 ระดับพหุสัมพันธ์ของตัวแปรอิสระที่ศึกษา

มี 3 ระดับ คือ

ระดับต่ำ ρ = 0.10 , 0.30

ระดับปานกลาง ρ = 0.50 , 0.70

ระดับสูง ρ = 0.90 , 0.99

โดย ρ คือ ความสัมพันธ์ระหว่าง x_1 กับ x_2 , x_1 กับ x_3 , x_2 กับ x_3

2) กรณีตัวแปรอิสระเท่ากับ 5 ระดับพหุสัมพันธ์ของตัวแปรอิสระที่ศึกษา มี 3 ระดับ คือ

ระดับต่ำ ρ = (0.10 , 0.10) , (0.30 , 0.30)

ระดับปานกลาง ρ = (0.50 , 0.50) ,(0.70 , 0.70)

ระดับสูง ρ = (0.90 , 0.90) ,(0.99 , 0.99)

โดย ρ ค่าแรกในวงเล็บ คือ ความสัมพันธ์ระหว่าง x_1 กับ x_2 , x_1 กับ x_3 , x_2 กับ x_3

และ ρ ค่าหลังในวงเล็บ คือ ความสัมพันธ์ระหว่าง x_4 กับ x_5

จ) ค่าสัมประสิทธิ์การถดถอยพหุที่ใช้ศึกษา คือ

ค่าเวกเตอร์เจาะจงซึ่งสอดคล้องกับค่าเจาะจงที่มีค่าน้อยที่สุดซึ่งทำให้ $E[L(R)^2] = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)]$ มีค่ามากที่สุด

4.2 ความคลาดเคลื่อนมีการแจกแจงปกติปลอมปน (Scale-Contaminated Normal Distribution)

ฟังก์ชันการแจกแจงอยู่ในรูปของ

$$F = (1-p)\,N(1,\sigma^2) + pN(1,c^2\sigma^2)$$

เมื่อ c คือ สเกลแฟคเตอร์(scale factor) ถ้าสเกลแฟคเตอร์มีค่าสูงจะทำให้เกิดค่าสังเกตที่ผิดปกติมีค่าสูงด้วย ในที่นี้จะใช้ c = 3 และ c = 10

และ 100p คือ เปอร์เซนต์การปลอมปน(percent of contaminate) ในกรณีนี้จะใช้ p = .05 และ .10

4.3 ความคลาดเคลื่อนมีการแจกแจงแบบเบ้

ผู้วิจัยจะศึกษาในกรณีการแจกแจงลอกนอร์มอล (Lognormal Distribution) โดยที่ฟังก์ชันความหนาแน่นดังกล่าวอยู่ในรูปของ

6

$$f(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2} \dfrac{(\ln(x) - \mu)^2}{\sigma^2} \right] & ; \; x > 0 \, , \; \sigma > 0 \\ \\ 0 & \text{อื่น ๆ} \end{cases}$$

เมื่อ μ และ σ^2 เป็นค่าเฉลี่ยและความแปรปรวนของ y ตามลำดับ โดยที่ y = ln(x) จะมีการแจกแจงปกติ ในที่นี้จะศึกษาโดยใช้ C.V. = 22% , 59% และ 100% ซึ่งจะสอดคล้องกับค่าความแปรปรวน σ^2 = 0.05 , 0.30 และ 0.70 ตามลำดับ และกำหนดค่า μ = 1 สาเหตุที่จะไม่เลือกค่า C.V. ที่ต่ำกว่านี้เพราะว่าถ้าค่า C.V. ต่ำกว่านี้กราฟของการแจกแจงจะสู่เข้าสู่การแจกแจงปกติ

5. เกณฑ์ที่ใช้พิจารณา

เกณฑ์ที่ใช้ในการพิจารณาจะแบ่งเป็น 2 กลุ่มใหญ่ คือ

1. เปรียบเทียบวิธีกำลังสองน้อยที่สุดกับวิธีริดจ์รีเกรสชัน และเปรียบเทียบวิธีกำลังสองน้อยที่สุดกับวิธีที่ใช้หลักการของริดจ์และสไตน์ โดยจะใช้ค่า PRR และ PRS ตามลำดับ ซึ่งจะคำนวณได้ดังนี้

$$PRR = \left[\frac{AMSE(OLS) - AMSE(RR)}{AMSE(OLS)} \right] \times 100$$

$$PRS = \left[\frac{AMSE(OLS) - AMSE(RS)}{AMSE(OLS)} \right] \times 100$$

สาเหตุที่ใช้เกณฑ์นี้เนื่องจากในทางทฤษฎีเราทราบว่าวิธีริดจ์รีเกรสชันและวิธีที่ใช้หลักการของริดจ์และสไตน์จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด

2. เปรียบเทียบวิธีริดจ์รีเกรสชันกับวิธีที่ใช้หลักการของริดจ์และสไตน์จะใช้ค่า RRS ซึ่งคำนวณได้ดังนี้

$$RRS = \left[\frac{AMSE(RS) - AMSE(RR)}{AMSE(RS)} \right] \times 100$$

สาเหตุที่เลือกใช้เกณฑ์นี้เนื่องจากผลการวิจัยโดยส่วนใหญ่เราทราบว่า วิธีริดจ์รีเกรสชันจะให้ค่าเฉลี่ยความคลาดเคลื่อนต่ำกว่าวิธีที่ใช้หลักการของริดจ์และสไตน์

6. วิธีดำเนินการวิจัย

1. สร้างข้อมูลตัวแปรอิสระโดยใช้เทคนิคมอนติคาโลตามระดับความสัมพันธ์ต่าง ๆ ที่กำหนดไว้ในขอบเขตการวิจัย โดยในแต่ละกรณีจะผลิตทั้งสิ้น 500 รอบ

2. สร้างข้อมูลตัวแปรตามจากตัวแปรอิสระที่สร้างขึ้น ณ ระดับพหุสัมพันธ์ที่กำหนดไว้ในขอบเขตการวิจัย และค่าความคลาดเคลื่อนตามการแจกแจงต่างๆ ที่กำหนดไว้

3. ประมาณค่าพารามิเตอร์ตามวิธีกำลังสองน้อยที่สุด วิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของริดจ์และสไตน์

4. คำนวณหาค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ของแต่ละวิธี

5. หาค่าเกณฑ์ที่ใช้ในการพิจารณา

ก) เปรียบเทียบวิธีกำลังสองน้อยที่สุดกับวิธีริดจ์รีเกรสชัน และเปรียบเทียบวิธีกำลังสองน้อยที่สุดกับวิธีที่ใช้หลักการของริดจ์และสไตน์ โดยใช้ค่า PRR และ PRS ตามลำดับ ซึ่งคำนวณได้ดังนี้

$$PRR = \left[\frac{AMSE(OLS) - AMSE(RR)}{AMSE(OLS)}\right] \times 100$$

$$PRS = \left[\frac{AMSE(OLS) - AMSE(RS)}{AMSE(OLS)}\right] \times 100$$

ข) เปรียบเทียบวิธีริดจ์รีเกรสชันกับวิธีที่ใช้หลักการของริดจ์และสไตน์ โดยใช้ค่า RRS ซึ่งคำนวณได้ดังนี้

$$RRS = \left[\frac{AMSE(RS) - AMSE(RR)}{AMSE(RS)}\right] \times 100$$

6. สรุปและอภิปรายผล

7. สัญลักษณ์ที่ใช้ในการวิจัย

1. AMSE(OLS) หมายถึง ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุเมื่อใช้วิธีกำลังสองน้อยที่สุด

2. AMSE(RR) หมายถึง ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุเมื่อใช้วิธีริดจ์รีเกรสชัน

3. AMSE(RS) หมายถึง ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุเมื่อใช้วิธีที่ใช้หลักการของริดจ์และสไตน์

4. PRR หมายถึง ร้อยละของอัตราส่วนผลต่างของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองระหว่างวิธีกำลังสองน้อยที่สุดและวิธีริดจ์รีเกรสชันเทียบกับวิธีกำลังสองน้อยที่สุด

5. PRS หมายถึง ร้อยละของอัตราส่วนผลต่างของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองระหว่างวิธีกำลังสองน้อยที่สุด และวิธีที่ใช้หลักการของริดจ์และสไตน์เทียบกับวิธีกำลังสองน้อยที่สุด

6. RRS หมายถึง ร้อยละของอัตราส่วนผลต่างของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองระหว่างวิธีที่ใช้หลักการของริดจ์และสไตน์และวิธีที่ใช้หลักการของริดจ์เทียบกับวิธีที่ใช้หลักการของริดจ์และสไตน์

8. ประโยชน์ที่คาดว่าจะได้รับ

ผลการศึกษาจะเป็นแนวทางในการเลือกใช้วิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุ ในกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์กัน เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ ปกติปลอมปน และแบบเบ้

9. สถิติที่ใช้ในการวิจัย

ในที่นี้จะกล่าวถึงวิธีการประมาณสัมประสิทธิ์การถดถอยพหุทั้ง 3 วิธี ซึ่งได้แก่ วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method (OLS)) วิธีริดจ์รีเกรสชัน (Ridge Regression Method (RR)) และวิธีที่ใช้หลักการของริดจ์และสไตน์ (Ridge and Stein Method (RS)) ซึ่งรายละเอียดต่าง ๆ มีดังนี้

9.1 การประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธี OLS

จากตัวแบบในสมการ (1) เราประมาณค่าเวกเตอร์พารามิเตอร์ β ด้วยวิธี OLS ซึ่งจะได้ค่าประมาณคือ $\hat{\beta} = (X'X)^{-1}X'y$ โดย $\hat{\beta}$ เป็นตัวประมาณที่ไม่เอนเอียงที่มีค่าเฉลี่ยความคลาดเคลื่อนต่ำที่สุดในบรรดาตัวประมาณที่ไม่เอนเอียง การประมาณค่า $\hat{\beta}$ ด้วยวิธี OLS มีข้อสมมติที่จำเป็นข้อหนึ่งคือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันซึ่งในทางปฏิบัติเป็นไปได้น้อยมาก ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กัน คือ มีสภาพที่ไม่เหมาะสม (ill-condition) การประมาณค่า β ด้วยวิธี OLS อาจจะไม่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ซึ่งการพิจารณาผลของตัวแปรอิสระที่มีพหุสัมพันธ์กันสามารถพิจารณาจากคุณสมบัติ 2 ประการ คือ

เมตริกซ์ความแปรปรวนและค่าความแปรปรวนร่วมของตัวประมาณ $\hat{\beta}$ และ ค่าเฉลี่ยของความแตกต่างกำลังสองระหว่าง $\hat{\beta}$ และ β ซึ่งสามารถเขียนให้อยู่ในรูปของ $X'X$ และ σ^2 ได้ดังนี้

(2) $$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

ให้ L_1 คือความแตกต่างระหว่าง $\hat{\beta}$ กับ β ดังนั้นค่าความแตกต่างกำลังสองระหว่าง $\hat{\beta}$ และ β มีค่าเป็น

$$L_1^2 = (\hat{\beta}-\beta)'(\hat{\beta}-\beta)$$

ค่าเฉลี่ยความแตกต่างกำลังสองระหว่าง $\hat{\beta}$ และ β คือ $E(L_1^2)$ ซึ่งมีค่าดังนี้

(3) $$E(L_1^2) = \sigma^2\text{trace}(X'X)^{-1}$$

ถ้า ε มีการแจกแจงปกติจะได้ว่า

(4) $$\text{Var}(L_1^2) = 2\sigma^4\,\text{trace}(X'X)^{-2}$$

เราเห็นได้ว่าจากสมการ (3) และ (4) จะอยู่ในรูปฟังก์ชันซึ่งเป็นผลบวกของสมาชิกในแนวทแยงมุมของเมตริกซ์ $X'X$ ($\text{trace}(X'X)$) ดังนั้นเพื่อความสะดวกในการเปรียบเทียบการประมาณค่า $\hat{\beta}$ จึงควรแปลงให้อยู่ในรูปฟังก์ชันของค่าเจาะจง (eigenvalue) ของเมตริกซ์ $X'X$ โดยใช้คุณสมบัติที่สำคัญข้อหนึ่งของค่าเจาะจง กล่าวคือ ถ้า λ_i เป็นค่าเจาะจงของเมตริกซ์ $X'X$ เมื่อ $i = 1,2,3,\ldots, p$ แล้ว $\sum_{i=1}^{p}\lambda_i = \text{trace}(X'X)$

สมมติให้ค่า λ_i เป็นค่าเจาะจงของเมตริกซ์ $X'X$ ซึ่ง λ_{max} คือค่าเจาะจงที่มีค่ามากที่สุดและ λ_{min} คือค่าเจาะจงที่มีค่าน้อยที่สุด ดังนั้น $(\lambda_{max} = \lambda_1) \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq (\lambda_p = \lambda_{min}) > 0$

จากสมการ (3) และ (4) เราสามารถเขียนในรูปฟังก์ชันของค่าเจาะจง ได้ดังนี้

$$E(L_1^2) = {}^{2}\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

และ $$\text{Var}(L_1^2) = 2\sigma^4\sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)^2$$

ดังนั้นในกรณีที่ตัวแปรอิสระมีสภาพที่ไม่เหมาะสม กล่าวคือ ค่าเจาะจงบางค่าของเมตริกซ์ $X'X$ จะมีค่าน้อยมาก ๆ ซึ่งมีผลทำให้ความแตกต่างระหว่าง $\hat{\beta}$ และ β มีค่ามาก นั่นคือ การประมาณค่า β ด้วยวิธี OLS จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่มีค่าสูงขึ้น ดังนั้นในกรณีที่เกิดปัญหาพหุสัมพันธ์ตัวประมาณ $\hat{\beta}$ จึงเป็นตัวประมาณที่ไม่เหมาะสม

9.2 การประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธี RR

Hoerl and Kennard (1970:55-67) ได้เสนอวิธี RR เพื่อแก้ปัญหาการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระซึ่งหลักการของตัวประมาณนี้คือ พยายามที่จะลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของการประมาณ $\hat{\beta}$ ให้ต่ำลง เมื่อพิจารณาจากค่าเจาะจงของเมตริกซ์ $X'X$ พบว่าในกรณีที่เกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระ ค่าเจาะจงบางค่าจะมีค่าน้อยมาก ๆ จึงทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของการประมาณ $\hat{\beta}$ มีค่าสูงขึ้น วิธี RR จึงแก้ปัญหานี้โดยการบวกค่าคงที่ที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุมของเมตริกซ์ $X'X$ เพื่อจะทำให้ได้ค่าเจาะจงที่สูงขึ้น โดยที่เราสามารถเขียนสมการปกติของตัวประมาณด้วยวิธี RR ได้เป็น

$$(X'X+kI)\hat{\beta}_R = X'y$$

ดังนั้นตัวประมาณ $\hat{\beta}_R$ จะอยู่ในรูปของ

(5) $$\hat{\beta}_R = (X'X+kI)^{-1}X'y \qquad ; \quad k>0$$

จากสมการ (5) การประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธี RR จะต้องกำหนดค่า k ซึ่งได้มีผู้เสนอวิธีในการประมาณค่า k หลายวิธีเช่น วิธีของ Hoerl Kennard and Baldwin (1975) วิธีของ TZE-SAN-LEE และวิธีของ McDonald Galarneau (1975) ซึ่งจากผลงานวิจัยของเจษฎาพร ขุทธนพิบูลย์ชัย(พ.ศ. 2533) ไม่สามารถสรุปได้แน่นอนว่าวิธีการประมาณค่า k วิธีการใดที่ให้ผลสรุปชัดเจน แต่จากผลงานวิจัยของจิราพุส พู่มนตรี(พ.ศ.2534) ได้มีการเสนอวิธีการประมาณค่า k โดยใช้วิธีการ Binary Search ซึ่งเป็นวิธีการหนึ่งในทางคอมพิวเตอร์ที่ใช้ในการค้นหาข้อมูลและผลสรุปที่ได้คือ วิธีการ Binary Search จะให้ผลในการประมาณค่า k ดีที่สุด แต่วิธีการนี้มีข้อจำกัดคือ ใช้กับข้อมูลที่มีลักษณะแบบไม่ต่อเนื่องและใช้ในการค้นหาข้อมูลที่มีลักษณะเรียงลำดับ ดังนั้นจึงต้องมีการเรียงข้อมูลจากน้อยไปหามากหรือจากมากไปหาน้อย จากข้อจำกัดดังกล่าวจึงทำให้วิธีการแบบ Binary Search ไม่เหมาะสมที่จะนำมาใช้ ผู้วิจัยจึงพยายามหาวิธีการค้นหาข้อมูลอื่นๆ ซึ่งพบว่าวิธีการค้นหาข้อมูลทางคอมพิวเตอร์ที่สามารถนำมาประยุกต์ใช้ได้คือ วิธีการค้นหาข้อมูลแบบลำดับ (Sequential Search) เพราะวิธีการนี้สามารถใช้ได้กับข้อมูลที่มีลักษณะทั้งแบบต่อเนื่องและแบบไม่ต่อเนื่องและข้อมูลที่จะค้นหาจะเรียงลำดับหรือไม่เรียงลำดับก็ได้ แต่ข้อเสียของวิธีการนี้คือ เราอาจใช้เวลาในการค้นหานานเมื่อข้อมูลที่ต้องการค้นหาอยู่ ณ ลำดับไกลๆ ซึ่งเมื่อพิจารณาจากคุณสมบัติของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุโดยทั่วไปพบว่าค่า k ที่เหมาะสมจะมีค่าไม่ห่างจากศูนย์มากนัก และ

ค่าของ k จะอยู่ในช่วง (0,1) จากคุณสมบัติของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยพหุลักษณะกราฟที่ได้จะมีจุดต่ำสุดเพียงจุดเดียวและไม่มีจุดวกกลับ ดังนั้นการพิจารณาที่จะใช้วิธีการค้นหาแบบลำดับจึงมีความเหมาะสมมากกว่า ผู้วิจัยจึงได้ใช้วิธีการในการประมาณค่า k คือวิธีการค้นหาแบบลำดับ (Sequential Search) โดยมีหลักการดังนี้

กำหนดให้ k_{opt} คือ ค่า k ที่ทำให้ $MSE(\hat{\beta}_R(k))$ มีค่าต่ำสุด

k_{int} คือ ค่า k ณ จุดเริ่มต้นในการค้นหาข้อมูล

k_{01} คือ ค่า k ณ จุดที่มีการเว้นช่วงห่างของค่า k เท่ากับ 0.01

d คือ ค่าคงที่ที่กำหนดขึ้นเพื่อเป็นช่วงห่างของค่า k

<u>ขั้นที่ 1</u> กำหนดให้ค่า $k_{int} = 0.0$ และ ค่า $d = 0.01$

<u>ขั้นที่ 2</u> คำนวณค่า $k_{01} = k_{int} + d$

<u>ขั้นที่ 3</u> คำนวณหา $MSE(\hat{\beta}_R(k_{int}))$ และหา $MSE(\hat{\beta}_R(k_{01}))$ แล้วทำการเปรียบเทียบ

ก) ถ้า $MSE(\hat{\beta}_R(k_{01})) > MSE(\hat{\beta}_R(k_{int}))$

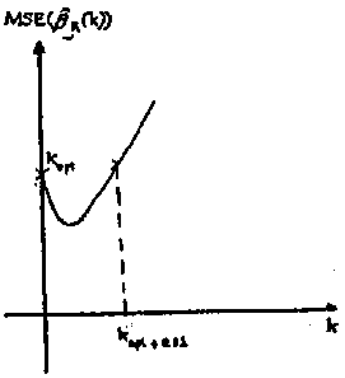
จะได้ว่า $k_{opt} = k_{int}$ ยุติการประมวลผล

ข) ถ้า $MSE(\hat{\beta}_R(k_{01})) \le MSE(\hat{\beta}_R(k_{int}))$ เราจะกำหนดให้

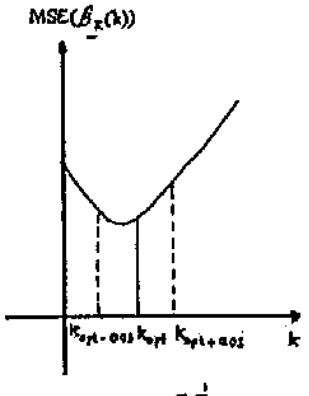
$k_{int} = k_{01}$

และทำการวิเคราะห์ขั้นที่ 2 ใหม่จนกว่าจะเข้ากรณี ก)

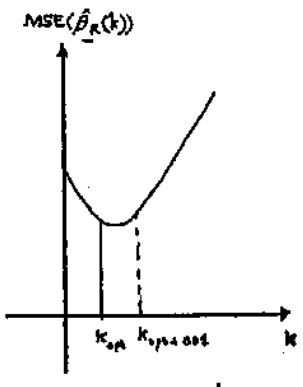
เนื่องจากลักษณะของ $MSE(\hat{\beta}_R(k))$ เป็นรูปโค้งหงาย การหาจุดที่ต่ำสุดจึงขึ้นอยู่กับการกำหนดช่วงห่างของค่า k ซึ่งจากการวิจัยพบว่าลักษณะความเป็นไปได้ของการจะพบจุดที่ต่ำสุดเมื่อช่วงห่างไม่ละเอียดพออาจแสดงได้ดังกรณีที่ 1-3



<u>กรณีที่ 1</u>



<u>กรณีที่ 2</u>



<u>กรณีที่ 3</u>

<u>กรณีที่ 1</u> ถ้าเรากำหนดช่วง d ห่างเกินไป จุดที่จะได้จุดต่ำสุดอาจจะเป็นจุดเริ่มแรกที่กำหนดคือ $k_{opt} = 0.0$ ซึ่งกรณีนี้ผู้วิจัยจะทำการแก้ปัญหาโดยให้มีการกำหนดช่วงห่างของการค้นหาให้ละเอียดขึ้นแล้วทำการค้นหาข้อมูลแบบลำดับจากจุดเริ่มต้น $k_{opt} = 0.0$ จนถึงจุดที่ $k_{opt} = 0.0 + 0.01$ ใหม่อีกครั้งโดยใช้ช่วงห่างของการค้นหาเป็น $d = 0.001$

<u>กรณีที่ 2</u> ถ้าจุดของค่า k ที่ค้นหาได้ตกอยู่หลังค่า k ที่เหมาะสมจริงๆ จึงจะเกิดกรณีนี้ขึ้น ผู้วิจัยได้แก้ปัญหาโดยการเพิ่มช่วงห่างของการค้นหาให้ละเอียดขึ้นและจะค้นหาข้อมูลทางด้านซ้ายของค่า k_{opt} อีกครั้งโดยจะค้นหาข้อมูลแบบลำดับจากจุด $k_{opt} - 0.01$ ไปจนถึงจุด k_{opt} โดยใช้ช่วงห่างของการค้นหาเป็น $d = 0.001$

<u>กรณีที่ 3</u> ถ้าจุดของค่า k ที่ค้นหาได้ตกอยู่ก่อนหน้าค่า k ที่เหมาะสมจริงๆ จึงจะเกิดกรณีนี้ ผู้วิจัยได้แก้ปัญหาโดยการเพิ่มช่วงห่างของการค้นหาให้ละเอียดขึ้นและจะค้นหาข้อมูลทางด้านขวาของค่า k_{opt} อีกครั้งโดยจะค้นหาข้อมูลแบบลำดับจากจุด k_{opt} ไปจนถึงจุด $k_{opt} + 0.01$ โดยใช้ช่วงห่างของการค้นหาเป็น $d = 0.001$

ดังนั้นผู้วิจัยจึงจะทำการค้นหาค่า k ที่เหมาะสมอีกครั้งโดยกำหนดช่วงความห่างของการค้นหาใหม่ คือ ให้ $d = 0.001$ โดยจะทำการตรวจสอบก่อนว่าค่า k_{opt} ที่ได้มีค่าเท่ากับศูนย์หรือไม่ ถ้ามีค่าเท่ากับศูนย์ก็คือกรณีที่ 1 ผู้วิจัยจะทำการค้นหาข้อมูลแบบลำดับเฉพาะทางด้านขวาของค่า k_{opt} แต่ถ้าค่า k_{opt} มีค่าไม่เท่ากับศูนย์ ผู้วิจัยจะทำการค้นหาข้อมูลแบบลำดับทางด้านซ้ายของค่า k_{opt} ก่อนถ้าไม่พบก็จะทำการค้นหาทางด้านขวาของค่า k_{opt} อีกครั้ง

เนื่องจาก ณ จุดที่ค่า $d = 0.001$ และ $d = 0.0001$ จะให้ค่า $MSE(\hat{\beta}_R(k))$ ที่มีค่าใกล้เคียงกัน ดังนั้นผู้วิจัยจึงไม่ทำการขยายช่วงค่า d จาก $d = 0.001$ เป็น $d = 0.0001$

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธี RR จะมีค่าเท่ากับ

$$\begin{aligned} MSE(\hat{\beta}_R) &= E(\hat{\beta}_R - \beta)^2 \\ &= \sigma^2 [\, \mathrm{trace}(X'X + kI)^{-1} - k\,\mathrm{trace}(X'X + kI)^{-2} \,] \\ &\quad + k^2 \beta'(X'X + kI)^{-2}\beta \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta'(X'X + kI)^{-2}\beta \end{aligned}$$

ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธี RR จะต้องเลือกค่า k ที่ทำให้ค่าความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์การถดถอยพหุมีค่าลดลง มากกว่าการเพิ่มขึ้นของ

ความเอนเอียงยกกำลังสอง จึงจะทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี RR มีค่าน้อยกว่าค่าเฉลี่ยความคลาดเคลื่อนกำลังสองจากวิธี OLS

9.3 การประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธี RS

Liu Keijian (1993:343-402) ได้เสนอวิธี RS ในการแก้ปัญหาการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระโดยใช้หลักการของตัวประมาณริดจ์ กล่าวคือการบวกค่าคงที่ค่าหนึ่งให้กับสมาชิกในแนวทแยงมุมของเมตริกซ์ $X'X$ แต่เนื่องจากตัวประมาณริดจ์มีปัญหาในการประมาณค่าพารามิเตอร์ k ดังนั้นตัวประมาณที่ใช้หลักการของริดจ์และสไตน์จึงแก้ปัญหาดังกล่าวโดยการบวกค่าคงที่ k ที่มีค่าเท่ากับหนึ่งให้กับสมาชิกทุกตัวในแนวทแยงมุมของเม-ตริกซ์ $X'X$ เนื่องจากตัวประมาณสไตน์เป็นตัวประมาณที่หาค่าพารามิเตอร์ง่ายเราจึงนำตัวประมาณดังกล่าวมาบวกกับเทอม $X'y$ เพราะฉะนั้นตัวประมาณอยู่ในรูปของ $\hat{\beta}_s = c\hat{\beta}$ เมื่อ $0 < c < 1$

การประมาณสัมประสิทธิ์การถดถอยพหุโดยวิธี RS จะได้ตัวประมาณที่อยู่ในรูปของ

$$\hat{\beta}_C = (X'X+I)^{-1}(X'y+c\hat{\beta}) \quad , \quad 0 < c < 1 \tag{6}$$

เราสามารถเขียน $\hat{\beta}_C$ ให้อยู่ในรูปฟังก์ชันของ $\hat{\beta}$ ซึ่งเราจะแทนค่า $X'y$ ที่ได้จากสมการปกติ $X'X\hat{\beta} = X'y$ ในสมการ (6) จะได้ว่าตัวประมาณดังกล่าวอยู่ในรูปของ

$$\begin{aligned}\hat{\beta}_C &= (X'X+I)^{-1}(X'X\hat{\beta}+c\hat{\beta}) \quad , \quad 0 < c < 1 \\ &= (X'X+I)^{-1}(X'X+cI)\hat{\beta} \\ &= C\beta\end{aligned}$$

เมื่อ $C = (X'X+I)^{-1}(X'X+cI)$ ถ้า $c = 1$ จะได้ $\hat{\beta}_C = \hat{\beta}$ และ $\hat{\beta}_C < \hat{\beta}$, $0 < c < 1$

เมื่อพิจารณาค่าคาดหวังของตัวประมาณที่ใช้หลักการของริดจ์และสไตน์จะได้ว่า

$$\begin{aligned}E(\hat{\beta}_C) &= E(C\hat{\beta}) \\ &= C\beta \\ &= (X'X+I)^{-1}(X'X+cI)\beta\end{aligned}$$

ความเอนเอียง(bias) ของ $\hat{\beta}_C = [E(\hat{\beta}_C) - \beta]$

$= [C\beta - \beta]$

(9) $$c = \sum_{i=1}^{p} \frac{\underset{\sim}{\alpha}_i^2 - \sigma^2}{(\lambda_i + 1)^2} \Big/ \sum_{i=1}^{p} \frac{\sigma^2 + \lambda_i \underset{\sim}{\alpha}_i^2}{\lambda_i (\lambda_i + 1)^2}$$

เมื่อแทนค่าตัวประมาณที่ไม่เอนเอียงของ $\underset{\sim}{\alpha}^2$ และ σ^2 ซึ่งมีค่าเท่ากับ $\hat{\underset{\sim}{\alpha}}^2 - \left(\frac{\hat{\sigma}^2}{\lambda_i}\right)$ และ $\hat{\sigma}^2$ ตามลำดับ ลงในสมการ (9) จะได้ว่าค่า c มีค่าเท่ากับ

(10) $$\hat{c}_{mm} = 1 - \hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\underset{\sim}{\alpha}}_i^2}{(\lambda_i + 1)^2}} \right]$$

จากสมการ (10) เราสามารถเขียนเป็นกรณีทั่วๆ ไปได้ คือ

$$\hat{c}_{mmh} = 1 - h\hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\underset{\sim}{\alpha}}_i^2}{(\lambda_i + 1)^2}} \right] \quad ; \quad h > 0$$

ค่าพารามิเตอร์ c สามารถใช้หลักการของการหาค่าพารามิเตอร์ k ได้ ซึ่ง Liu Keijian (1993:343-402) ได้ทำการศึกษาโดยใช้วิธีการทำซ้ำกับหลักวิธีการของ Hoerl and Kennard , GCV(Generalized Cross Validation) และวิธีที่ใช้เกณฑ์ C_L (C_L -Criterion) ซึ่งผลสรุปที่ได้ คือ วิธีการทำซ้ำของ Hoerl and Kennard จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยที่มีค่าต่ำสุด ดังนั้นในการวิจัยครั้งนี้จะใช้ค่า c ที่ได้จากการทำซ้ำของ Hoerl and Kennard ซึ่งมีขั้นตอนการทำซ้ำคือ

<u>ขั้นที่ 1</u> $\hat{\underset{\sim}{\alpha}}$ $$c_0 = 1 - \hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\underset{\sim}{\alpha}}_i^2}{(\lambda_i + 1)^2}} \right]$$

<u>ขั้นที่ 2</u> $\hat{\underset{\sim}{\alpha}}_c(c_0)$ $$c_1 = 1 - \hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\hat{\underset{\sim}{\alpha}}_{c_i}^2(c_0)}{(\lambda_i + 1)^2}} \right]$$

ขั้นที่ 3 $\underset{\sim}{\hat{\alpha}}_C(c_1)$ $c_2 = 1-\hat{\sigma}^2\left[\dfrac{\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}}{\sum_{i=1}^{p}\frac{\hat{\alpha}_{ci}^{\ 2}(c_1)}{(\lambda_i+1)^2}}\right]$

ผู้วิจัยจะยุติการทำซ้ำถ้า $\dfrac{(c_j-c_{j+1})}{c_j} \le 20\,T^{-1.3}$ ซึ่ง $T=\dfrac{\text{trace}(X'X)^{-1}}{p}$

การศึกษาคุณสมบัติของตัวประมาณที่ใช้หลักการของริดจ์และสไตน์มีทฤษฎีต่างๆ ที่เกี่ยวข้องดังนี้

ทฤษฎีที่ 2.1 ถ้า $0<c<1$ จะได้ว่า $MSE(\underset{\sim}{\hat{\beta}}_C) < MSE(\underset{\sim}{\hat{\beta}})$

ทฤษฎีที่ 2.2 สมมติว่า $\underset{\sim}{\varepsilon} \sim N(0,\sigma^2 I)$ ค่า h ที่เหมาะสมคือ

$0< h < \dfrac{2(n-p)}{n-p+2}\left[1-\dfrac{2}{\sum_{i=1}^{p}\frac{\lambda_p(\lambda_p+1)}{\lambda_i(\lambda_i+1)}}\right]$ สำหรับทุกค่าของ β_i และ σ^2 จะทำให้ได้ว่า

$MSE(\underset{\sim}{\hat{\beta}}_C(\hat{c}_{mmh})) < MSE(\underset{\sim}{\hat{\beta}})$

บทแทรก 2.2.1 เมื่อ $\dfrac{2(n-p)}{n-p+2}\left[1-\dfrac{2}{\sum_{i=1}^{p}\frac{\lambda_p(\lambda_p+1)}{\lambda_i(\lambda_i+1)}}\right] >$ สำหรับทุกค่า $\underset{\sim}{\beta}$ และ σ^2

จะได้ว่า $MSE(\underset{\sim}{\hat{\beta}}_C(\hat{c}_{mmh})) < MSE(\underset{\sim}{\hat{\beta}})$

10. สรุปผลการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง ได้ศึกษาในสถานการณ์ต่างๆ ดังนี้

ก) ความคลาดเคลื่อนมีการแจกแจงปกติ ปกติปลอมปน และลอกนอร์มอล

ข) ขนาดตัวอย่าง(n) เท่ากับ 30 , 50 และ 100

ค) จำนวนตัวแปรอิสระที่ศึกษามีดังนี้

1) จำนวนตัวแปรอิสระเท่ากับ 3 จำแนกตามระดับพหุสัมพันธ์ได้ดังนี้

-ระดับต่ำเท่ากับ 0.10 และ 0.30

-ระดับปานกลางเท่ากับ 0.50 และ 0.70

-ระดับสูงเท่ากับ 0.90 และ 0.99

2) จำนวนตัวแปรอิสระเท่ากับ 5 จำแนกตามระดับพหุสัมพันธ์ได้ดังนี้

-ระดับต่ำเท่ากับ (0.10,0.10) และ (0.30,0.30)

-ระดับปานกลางเท่ากับ (0.50,0.50) และ (0.70,0.70)

-ระดับสูงเท่ากับ (0.90,0.90) และ (0.99,0.99)

ตารางที่ 1.1 สรุปสถานการณ์ที่วิธี OLS สามารถนำไปใช้ได้

การแจกแจงของค่าความคลาดเคลื่อน	จำนวนตัวแปรอิสระ	ระดับความสัมพันธ์	สถานการณ์
ปกติ	3	ระดับต่ำและระดับปานกลาง	ทุกระดับของ σ และ n
	5	ระดับต่ำและระดับปานกลาง	ทุกระดับของ σ และ n
ปกติปลอมปน	3	ระดับต่ำและระดับปานกลาง	c = 3 , p = 5,10 ทุกระดับของ σ และ n c = 10 , p = 5,10 n = 100 ทุกระดับของ σ
	5	ระดับต่ำ	c = 3 , p = 5,10 ทุกระดับของ σ และ n c = 10 , p = 5,10 n = 100 ทุกระดับของ σ
ลอกนอร์มอล	3	ระดับต่ำ	n = 100 ทุกระดับของσ^2

ตารางที่ 1.2 สรุปสถานการณ์ที่วิธี RR และ RS สามารถนำไปใช้ได้

การแจกแจงของค่าความคลาดเคลื่อน	จำนวนตัวแปรอิสระ	วิธีริดจ์รีเกรสชัน(RR)	วิธีที่ใช้หลักการของริดจ์และสไตน์(RS)
	3	-ระดับต่ำและระดับปาน	-ระดับต่ำและระดับ

การแจกแจงของค่าความคลาดเคลื่อน	จำนวนตัวแปรอิสระ	วิธีริดจ์รีเกรสชัน(RR)	วิธีที่ใช้หลักการของริดจ์และสไตน์(RS)
ปกติ		กลาง(0.50) เมื่อ σ=0.05 ,0.10 และ n = 30, 50 เมื่อ σ = 0.15 สำหรับทุกขนาดของ n -ระดับปานกลาง(0.70)และระดับสูง ทุกระดับของσ และ n	ปานกลาง(0.50) เมื่อ σ = 0.05 และ n = 100
	5	-ระดับต่ำและระดับปานกลาง เมื่อ σ = 0.05 และ n = 30,50 เมื่อ σ = 0.10 ,0.15 ทุกขนาดของ n -ระดับสูง ทุกระดับของ σ และ n	-ระดับต่ำและระดับปานกลาง เมื่อ σ = 0.05 และ n = 100
	3	-ระดับต่ำและระดับปานกลาง(0.50) ทุกระดับของ c, p,σ และ n ยกเว้น c = 3, p=5,10 และ σ=0.05 n=50, 100 -ระดับปานกลาง(0.70) และระดับสูง ทุกระดับของ c, p, σ และ n	-ระดับต่ำและระดับปานกลาง(0.50) c=3, p=5, 10, σ=0.05 และ n=50, 100

การแจกแจงของค่าความคลาดเคลื่อน	จำนวนตัวแปรอิสระ	วิธีริดจ์รีเกรสชัน(RR)	วิธีที่ใช้หลักการของริดจ์และสไตน์(RS)
ปกติปลอมปน	5	-ระดับต่ำและระดับปานกลาง(0.50) ทุกระดับของ c, p,σ และ n ยกเว้น c = 3, p=5,10 และ σ=0.05 n=100 -ระดับปานกลาง(0.70) และระดับสูง ทุกระดับของ c, p, σ และ n	-ระดับต่ำและระดับปานกลาง(0.50) c=3, p=5, 10, σ=0.05 และ n=100
ลอกนอร์มอล	3 และ 5	-ทุกระดับความสัมพันธ์ σ^2 และ n	

หมายเหตุ ค่า σ หมายถึง ค่าการกระจายของความคลาดเคลื่อนที่ศึกษา

ค่า c, p หมายถึง สเกลแฟคเตอร์และเปอร์เซนต์การปลอมปนเมื่อความคลาดเคลื่อนมีการแจกแจงปกติปลอมปน ตามลำดับ

จากตารางที่ได้สรุปแล้วข้างต้นอาจทำให้มีความยุ่งยากในการตัดสินใจเลือกวิธีการที่จะนำไปใช้เนื่องจากในทางปฏิบัติเมื่อเราได้ข้อมูลมาชุดหนึ่งอาจจะยังไม่สามารถทราบถึงการแจกแจงของความคลาดเคลื่อน ระดับความสัมพันธ์ของตัวแปรอิสระ หรือส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนจนกระทั่งต้องนำข้อมูลนั้นไปวิเคราะห์ก่อน ดังนั้นเพื่อให้เหมาะสมในการนำไปใช้งานผู้วิจัยจึงขอทำการสรุปถึงวิธีการภายใต้สถานการณ์ที่ทราบเพียงจำนวนตัวแปรอิสระและขนาดตัวอย่างเพราะข้อมูลทั้งสองเป็นข้อมูลพื้นฐานที่จะทราบได้เบื้องต้นซึ่งผลสรุปอยู่ในตารางข้างล่างนี้

จำนวนตัวแปรอิสระ	ขนาดตัวอย่าง		
	30	50	100
3	RR	RR	OLS , RS
5	RR	RR	RS